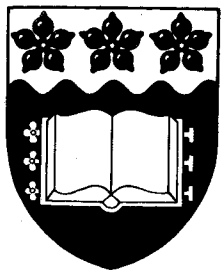
1990

Sample Size and the Strength of Evidence: A Bayesian Interpretation of Binomial Tests of the Information Content of Qualified Audit Reports

D. J. Johnstone
University of Wollongong

UNIVERSITY OF WOLLONGONG

DEPARTMENT OF ACCOUNTANCY



SAMPLE SIZE AND THE STRENGTH OF EVIDENCE

A Bayesian Interpretation of Binomial Tests of the Information Content of Qualified Audit Reports

by

David J. Johnstone

January 1990

WORKING PAPER NO. 13

This paper was presented at the Doctoral Consortium
on 7 July 1989, organised by
The Department of Accountancy, The University of Wollongong

SAMPLE SIZE AND THE STRENGTH OF EVIDENCE:

A Bayesian Interpretation of Binomial Tests of the Information Content of Qualified Audit Reports

D.J. Johnstone*

Abstract

Lindley (1957) demonstrated that from a Bayesian standpoint a given level of statistical significance P, carries less evidence against the null hypothesis H_0 the larger (more powerful) the test. Moreover, if the sample is sufficiently large, a result significant on H_0 at 5% or lower may represent strong evidence in support of H_0, not against it. Contrary to Lindley's argument, a great many applied researchers, trained exclusively in orthodox statistics, feel intuitively that to "reject" the null hypothesis H_0 at (say) α=5% is more convincing evidence, *ceteris paribus*, against H_0 the larger the sample. This is a consistent finding of surveys in empirical psychology. Similarly, in accounting, see Burgstahler (1987). In econometrics, "Lindley's paradox" (as it has become known statistics) has been explained in well known books by Zellner (1971), Leamer (1978) and Judge *et al.* (1982), but is not widely appreciated. The objective of this paper is to reiterate the Bayesian argument in an applied context familiar to empirical researchers in accounting.

*Graduate School of Management and Public Policy
University of Sydney
2006, Australia

November, 1989

1. Introduction

Surveys in psychology, including most recently Nelson *et al.* (1986), suggest that the great majority of researchers who use statistical tests of significance believe intuitively that the null hypothesis is discredited more convincingly by a larger sample (more powerful test) at the same level of significance. In accounting, the evidential interpretation of tests is not often discussed, not at least in published research. But in a methodological paper written for research students, Burgstahler takes apparently the same position as those researchers surveyed in psychology:

> For more powerful tests, there is a lower degree of belief in the null (and greater belief in the alternative) as a result of observation of a significant statistic. (Burgstahler, 1987, p.207)

Burgstahler claims that a "significant" result represents stronger evidence the higher the power of the test.[1] Consistent with this conclusion, the suggestion by some researchers (Burgstahler gives references) that a given level of significance entails greater evidence against the null hypothesis when the sample is small (power is low) than when the sample is very large (power is high) is expressly denied:

> ...researchers may even mistakenly assert that a significant result from a low-power test is more convincing evidence against the null than a significant result from a high-power test because a more extreme test statistic is required to attain significance for a low-power test. (Burgstahler, 1987, p.207)

Notwithstanding Burgstahler's anathema, it has been noted among theoretical statisticians at least since Lindley (1957) that under fairly general conditions the posterior probability of the null hypothesis, given a fixed level of significance P, increases (for a

given prior) with the sample size *n*. Thus, a "significant" result is more compelling if the sample size is small than large. This is known as "Lindley's (or Jeffreys') paradox". References include Jeffreys (1939, pp.359-360), Lindley (1957, pp.187-189), Pratt (1961, pp.165-166; 1976, p.782; 1977, pp.63-67), Edwards *et al.* (1963, pp.221-231), Zellner (1971, pp.303-306; 1984, pp.276-279), Rosenkrantz (1973, p.314; 1977, p.208), Cox and Hinkley (1974, pp.395-397), Basu (1975, pp.43-47), Leamer (1978, pp.104-105), Good (1980, pp.307-309; 1981, pp.163-164; 1982, p.342; 1983, pp.312-314; 1984, pp.300-302; 1985, p.260), Hill (1982, p.345), Judge *et al.* (1982, p.101), Berger (1985, p.156), DeGroot (1986, pp.449-450), Royall (1986, pp.313-345), Johnstone (1986, pp.493-494; 1989a, p.5; 1989c), and Berger and Sellke (1987, p.112). The passages quoted below are clear and authoritative:

> Prior beliefs about the null hypothesis aside, bare significance at the 5% level does not contradict the null hypothesis equally as the statistical problem varies or as the sample size varies in a given problem (for instance, testing $\rho=.5$ against $\rho=.6$ with 10 or 1000 binomial observations). Consider also a most powerful level 5% test for a simple hypothesis against a simple alternative having power 99%: if the distributions are continuous, a result which is just significant can hardly be said to favor the alternative, since it is also just significant at level 1% when the hypotheses are reversed. In fact, the more powerful the test, the more a just significant result favors the null hypothesis. (Pratt, 1961, pp.165-166)

> ...the interpretation to be placed on the phrase "significant at 5%" depends on the sample size: it is more indicative of the falsity of the null hypothesis with a small sample than with a large one. (Lindley and Scott, 1984, p.3).

> ...as the sample size increases in testing precise hypotheses, a given *P* value provides less and less real evidence against the null. (Berger and Sellke, 1987, p.136)

2. An Example

Dodd *et al.* (1984) examine the security returns of firms with qualified financial reports. Abnormal returns are measured over 3 and 5 day periods around the day qualification is made known. The authors model the return process as:

$$x \sim \text{Binomial } (n,\theta)$$

where x is the number of firms observed with negative abnormal returns, n is the total number of firms observed (the sample size), and θ is the probability with which firms with qualified financial reports yield negative abnormal returns. The hypothesis tested, H_O: $\theta=.5$, is that 50% of firms with qualified reports earn negative abnormal returns around the time of qualification. No alternative hypothesis is specified, but the test used is one-(right)-tailed. Thus, the alternative implied is H_A: $\theta>.5$.[2] Two subcategories of firms, those with qualifications attributed to (i) litigation and (ii) asset realizing,[3] exhibit statistically significant (at $\alpha=5\%$) proportions of negative abnormal returns over the 3 day observation interval. The results of these two tests are as follows:

Test (i) *litigation firms* $n = 130$

$x = 76$

P-level = .0325

Test (ii) *asset realizing firms* $n = 93$

$x = 56$

P-level = .0307

Both P-levels are about 3%. On a classical inferential (Fisherian) interpretation, the strength of evidence against H_O: $\theta=.5$ is the

same for both tests.[4] By comparison, on a Bayesian (probabilistic) interpretation, the evidence against H_o is stronger for the smaller (less powerful) test than for the test with the larger sample size. That is, there is greater evidence that θ exceeds 50% for asset realizing firms than for litigation firms. The mathematical basis for this inference is that the likelihood ratio, $Pr\{x|H_o\}/Pr\{x|H_a\}$, is lower for test (ii) than test (i) over the range of all point alternatives H_a: $\theta=\theta_a$, where $\theta_a>.58$. In the interval $.5<\theta_a\leq.58$, the likelihood ratios of test results (i) and (ii) are approximately equal, but as θ_a increases, the ratio for the result in test (ii) becomes much lower than that for the result in test (i).[5] Therefore, unless the prior probability of H_A: $\theta>.5$ is nearly all lumped in the range $.5<\theta_a\leq.58$, the posterior probability of H_o is less for test (ii) than test (i).[6] For prior distributions on H_A not concentrated in the range $.5<\theta_a\leq.58$, the posterior probability of the null hypothesis H_o is (much) lower on test (ii), the test with the smaller sample size, than for the larger test. That is, H_o is discredited more by the smaller (less powerful) test than by the test with the larger sample.

Test (i), which is the larger (more powerful) test, discredits H_o more strongly than test (ii) *only if the mass of prior probability on H_A is concentrated in the interval* $.5<\theta_a\leq.58$.[7] Hence, a reader who interprets the result in test (i) as stronger evidence against H_o than test (ii) assumes implicitly a prior distribution in this class.[8] Specifically, an inference (belief) based on the results of tests (i) and (ii) that H_o is less probable for litigation firms than asset realizing firms implies (if one accepts Bayes' theorem) the prior belief that if H_o: $\theta=.5$ is not true then θ is almost certainly in the interval $.5<\theta_a\leq.58$.[9]

For more diffuse prior distributions, even those with a relatively high probability on θ in the range $.5 < \theta_a \leq .58$, test (ii) provides the stronger evidence. For example, assume the prior probability distribution $\Pr\{H_o\}=\Pr\{H_A\}=.5$, where $\Pr\{H_A\}$ is distributed on H_o: $\theta>.5$ according to $g(\theta) \propto \theta^2(1-\theta)^4$.[10] This distribution gives both H_o and H_A prior probability .5,[11] which is necessary if the competing hypotheses H_o and H_A are to be evaluated "objectively" (impartially).[12] The prior density on $\theta>.5$, $g(\theta)$, represented by Figure 1 below, is proportional to a beta distribution with parameters a=3 and b=5 (note that the distribution shown has not been normalized). These parameter values are chosen such that $g(\theta)$ is maximum in the range (.5,1] at θ=.5 and tails off toward zero as θ approaches one.

[Figure 1 about here]

The posterior probability of H_o, $\Pr\{H_o|x\}$, based on the prior probability distribution described above, is given as follows:

$$\Pr\{H_o|x\} = \left[1 + \frac{\Pr\{H_A\}}{\Pr\{H_o\}} \frac{\Pr\{x|H_A\}}{\Pr\{x|H_o\}} \right]^{-1}$$

$$= \left[1 + \frac{1}{B} \right]^{-1}$$

$$\text{where } B = \frac{(.5)^n \int_{.5}^{1} \theta^2 (1-\theta)^4 \, d\theta}{\int_{.5}^{1} \theta^{x+2} (1-\theta)^{n-x+4} \, d\theta} \qquad \ldots\ldots(1)$$

is the weighted likelihood ratio or "Bayes' factor" in favor of the null hypothesis H_O against the alternative H_A. Note that for the purpose of revising probability assessments using Bayes' theorem, the likelihood ratio is an exhaustive summary of the information contained in the sample concerning the unknown parameter.

From equation (1), the posterior probability of H_O, $\Pr\{H_O|x\}$, is .2379 for test (i) and .2130 for test (ii).[13] Thus, assuming the prior density function $g(\theta)$ proportional to $\theta^2(1-\theta)^4$, the null hypothesis is discredited more strongly by test (ii) than test (i). There is not much difference between the two results, but if the prior on H_A is distributed more evenly toward larger values of θ, the difference between the Bayesian results on tests (i) and (ii) becomes greater. For example, if the prior probability on H_A, $\Pr\{H_A\}=.5$, is distributed uniformly, *i.e.*, if the density function $g(\theta)$ is constant on (.5,1], the posterior probability of H_O, $\Pr\{H_O|x\}$, is given by:

$$\Pr\{H_O|x\} = \left[1 + \frac{1}{B} \right]^{-1}$$

$$\text{where } B = \frac{(.5)^n}{2\int_{.5}^{1} \theta^x (1-\theta)^{n-x}\, d\theta} \qquad .....(2)$$

which equals .4236 for test (i) and .3652 for test (ii).

Note that for priors over H_A concentrated on values of θ around .7 or higher, the posterior probability of H_O is lower on test (ii) than test (i), but neither posterior is small. For example, if the prior

distribution gives simple hypotheses H_0: $\theta=.5$ and H_a: $\theta=\theta_a$, where $\theta_a=.7(.75)[.77]$, each probability .5 (other hypotheses are excluded *a priori*), the probability of the hypothesis H_0 is .88(.999)[.9999] on test (i), and .51(.95)[.99] on the result in test (ii). This is a simple example of Lindley's paradox, whereby results which are "statistically significant" in orthodox terms entail evidence *in favor* of the null hypothesis, not against it.

Lindley's paradox is apparent in both tests (i) and (ii) for prior distributions on H_A concentrated around $\theta \geq .7$. For distributions with more weight in the interval $.5<\theta<.7$, inconsistency between the orthodox and Bayesian inferences appears at larger sample sizes than those in the Dodd *et al.* study. If the sample size n is very large, the conventional and Bayesian inferences based on a "significant" result, $\{x$: P-level=$\alpha\}$, where α=5% say, are generally not consistent. For example, in a hypothesis test of H_0: $\theta=.5$ against H_A: $\theta>.5$, given the prior probability distribution presumed in equation (2) above, the result x=527(1037) in n=1000(2000) trials has (one-sided) P-level equal to .05, but gives Bayesian posterior probability $Pr\{H_0|x\}=.7546(.8269)$. Lindley (1957, p.190) notes that for more moderate sample sizes, such as those in tests (i) and (ii) compared in this paper, the conventional and Bayesian conclusions may often be consistent (although not often the same; the Bayesian inference is expressed as a probability on H_0 or on an interval around the null value of θ, whereas the orthodox inference is typically something less precise, such as a statement of the form "the null hypothesis is probably false", or "there is strong evidence against the null hypothesis"). Care should be taken, however, when the sample size is very large. See Johnstone (1989c) on the interpretation of large sample tests.

3. Other Dodd *et al.* Results

In this section, the Dodd *et al.* results of binomial tests for the entire sample (n=283) over 3 and 5 day observation intervals are examined. The Bayesian inferences based on these results are considered in relation to the orthodox P-values reported by the authors. For comparison of the Bayesian and orthodox results based on various other Dodd *et al.* observations, see Appendix A.

Table 1 below is taken from the Dodd *et al.* paper (Table 6, p.24) but with the addition of Bayesian posterior probabilities corresponding to the P-levels stated. These probabilities are based, for convenience, on the prior distribution $Pr\{H_O\}=Pr\{H_A\}=.5$, where $Pr\{H_A\}$ is distributed uniformly on $.5<\theta\leq 1$, as per equation (2) above. The sensitivity of the Bayesian results to the priors on which these are calculated is discussed in footnotes below.

Table 1

Excerpt from Dodd *et al.* Table 6

	All firms	
Event window	-1 to +1	-2 to +2
n = # of firms	283	283
x = # < 0	165	153
p = % < 0	58	54
P-value (binomial test)[a]	.0031	.0954
Bayesian $Pr\{H_O \mid x\}$[b]	.1194	.7435

[a] H_O: $\theta = .5$ against H_A: $\theta > .5$.

[b] Based on the prior density function $g(\theta)$ constant over $(.5, 1]$.

(*a*) *The -1 to +1 Result*. Here the orthodox and Bayesian inferences are consistent. The orthodox inference based on a P-level of .0031 is that there is strong evidence against the hypothesis tested; *either* the hypothesis H_0 is false *or* an event of very low probability (P-level) has occurred, hence it may be inferred that H_0 is probably false or in some sense strongly discredited (this is classical Fisherian logic). Consistent with the orthodox result, the Bayesian posterior probability of the null hypothesis is only .1194 (down from a prior of .5).[14] It is apparent, therefore, that if the P-level observed is low enough (*i.e.*, if the test statistic observed is discrepant enough with H_0), it does not matter that the sample is moderately large, the posterior probability of H_0 is low nonetheless. Generally, an extremely low P-value represents strong evidence against H_0 even if the sample is fairly large. Ultimately, however, Lindley's paradox takes over; if n is sufficiently large, even the smallest P-values support H_0. Note, for example, the results conditioned on P-levels .001 and .0001 tabled in Appendix B below.

(*b*) *The -2 to +2 Result*. If the sample size is large, a result just significant at say α=.05 or α=.01 may support the null hypothesis. For results "not significant" at these conventional partitions (*i.e.*, P-levels > .05), the sample size need not be very large before H_0 is (strongly) supported.[15] In the case of the *Dodd et al.* result for days -2 to +2, a P-level of .0954 with a sample of n=283 increases the probability of H_0 from .5 to .7435.[16] In this case, the Bayesian and conventional inferences are logically inconsistent. A conventional interpretation of the P-level .0966 is that there is neither (strong) evidence for or against H_0, which would suggest, contrary to the Bayesian inference, that the probability of H_0 is not (much) changed.

Note that if the P-level observed had been higher, and or the sample size larger, the Bayesian posterior probability of H_o would have increased more substantially. Suppose, for example, that the sample size had been n=500(1000)[5000]. In this case, assuming the same P-level .0954, the probability of H_o, given by equation (2), would have increased from .5 to .7999(.8521)[.9305]. It is clearly important, therefore, to interpret orthodox P-values very carefully. In particular, the sample size must be accounted for in conjunction with the P-level reported. This is explained in just a few orthodox textbooks; notably, Cox and Hinkley (1974, pp. 397-398), and Kendall and Stuart (1979, p.197). See also Hodges and Lehmann (1954), and Appendix B of this paper.

(c) *Excluding the "Busy" Period.* The classical interpretation of statistical significance tests (due to R. A. Fisher) makes no reference to the sample size (see footnote 4). It is not usual, therefore, that researchers condition the inferential interpretation of observed levels of significance on the size of the sample observed. Consider, for example, the paragraph below from the paper by Dodd *et al.*:

> The number of observations in the full sample falls from 283 to 210 when the 'busy' period [data for firms filing around March 30] is eliminated. ..The proportion of firms with negative abnormal returns in the period -1 to +1 increases from 58% to 61% ..which decreases the *P* value for the binomial test from 0.0031 to 0.0006. The proportion of negative returns in the asset realizing group in the same window rises to 67% from 60%, and the *P* value for the binomial test decreases from 0.0307 to 0.0049. Other *P* values are essentially unaltered. (p.28)

In this passage, the authors report changes in P-levels, but give no relevance to the corresponding changes in sample size. A lower

P-level is interesting in itself, but more so in conjunction with a smaller sample. In particular:

(i) the decrease in the P-level for all firms (-1 to +1) from .0031 to .0006 when the "busy" period is excluded strengthens the evidence against H_O: θ=.5; not only is the P-level smaller but so is the sample size.[17] The probability of H_O given by equation (2) falls from .1194 (n=283, P-level=.0031) to .0230 (n=210, P-level=.0006). Similarly, $Pr\{H_O|x\}$, calculated as per equation (1), falls from .0536 to .0118.

(ii) the decrease in the P-level for asset realizing firms (-1 to +1) cannot be interpreted immediately because the reduced sample size is not reported. It would appear, however, that the new sample size is n=66; a result x=44 in n=66 observations gives p=67% and P-level .0049 which is as reported by the authors. If n=66 is the size of the reduced sample, the probability of H_O: θ=.5, given a uniform prior on H_A, falls from .3652 (n=93, P-level=.0307) to .0766 (n=66, P-level =.0049). For the prior assumed in equation (1), the posterior probability of H_O falls from .2130 to .0561.

(iii) if firms filing around March 30 can be excluded from the sample, there is generally an increase in the evidence against the null hypothesis H_O: θ=.5. It is reported that most P-values are no lower than when the "busy" observations are included, however the sample sizes are lower. Generally, therefore, so is the Bayesian posterior probability, $Pr\{H_O|x\}$.

4. Conclusion

The result in a test of statistical significance should be interpreted with respect to the sample size.

> 5% in to-day's small sample does not mean the same as 5% in to-morrows large one. (Lindley, 1957, p.189)

In general, a fixed level of significance P represents less evidence against the null hypothesis as the sample size n increases.[18] Moreover, if n is large, a result which is significant at 5% or lower may often represent evidence in favor of the null hypothesis, not against it. To account for this logical "paradox", it is recommended in advanced orthodox textbooks (references are provided above) that the "critical" level of significance α be made more stringent (smaller) than usual when the sample size is large. If α is made (very) small when n is (very) large, then presumably the null hypothesis H_0 will not be "rejected" when the evidence, on a Bayesian interpretation, is that H_0 is (approximately) true.[19] Thus, by setting α lower than normal when n is large, an orthodox statistician can come to conclusions not inconsistent with those of a Bayesian.

There is, however, no theoretical basis on which to fix α with respect to n. In a decision theory approach, the test characteristic α (defined as the probability of a type I error) is determined by reference to the loss function, but if the test is to be interpreted *inferentially* (*e.g.*, "on the basis of the sample observed, the null hypothesis H_0 is highly improbable") rather than as a *decision* between alternate courses of action (*e.g.*, accept or reject the account balance stated by the company audited) there is no particular decision contemplated and hence no loss function to which to refer. It is not clear, therefore, without explicit Bayesian calculations, how to interpret a given level of significance in terms of evidence. The sample size is important, but there is no formal relationship by which P-levels can be calibrated Bayesianly [Berger and Delampady

(1987) pp.327-328; Berger and Sellke (1987) pp.135-136, 138]. This is of great concern to those researchers who employ orthodox techniques, particularly tests of significance, yet think of themselves, at least in principle, as Bayesian; *cf.*, Burgstahler (1987, p.204).

An orthodox P-level is a summary of data, but not in itself an (inductive) inference. Inference is what one makes of that P-level. The logical inference based on a reported P-level depends generally on whether the logic applied is orthodox (Fisherian) or Bayesian. Recently, statistical theorists have compared the Bayesian and conventional interpretations of orthodox P-levels over a broad class of statistical tests. Note especially the papers of Berger and Sellke (1987), Berger and Delampady (1987), and Casella and Berger (1987) (not the same Berger) in recent issues of the journals of the American Statistical Association (*JASA* and *Statistical Science*). In accounting, the paper by Burgstahler (1987) published in the *Accounting Review* takes ostensibly a Bayesian approach but comes erroneously to "orthodox" (contra-Bayesian) conclusions (note again footnote 1). Researchers who accept Burgstahler's argument may be under the impression that to be orthodox is in effect, or ideally, to be Bayesian. It is important, therefore, that empirical researchers in accounting, who may not have been trained in Bayesian statistics, appreciate that the Bayesian and conventional inferences predicated on the same P-level are not necessarily, or in general, the same, especially if the sample size is very large.

In particular, Bayesian theorists maintain that the larger the sample size, the weaker, generally, the evidence (against the null hypothesis) entailed by a given level of significance. This is of

particular concern in relation to securities market based research in accounting and finance where some samples are immensely large [Ball and Foster (1982) pp.186-187]. In studies such as these, results only just significant at levels such as 5% may more support the null hypothesis than discredit it. It may be useful, therefore, as in this paper, to re-examine previously published results using Bayesian instead of orthodox techniques. A great many published conclusions will be strongly confirmed (strengthened), particularly if the observed P-level is very low or the sample size not very large. However, more importantly perhaps, established results based on very large samples where the P-level is not extremely low may in some cases be qualitatively incompatible with Bayesian conclusions based on the same experiment. It is not logical (Bayesian) to argue that merely because the observed level of significance is say 5% or less, the null hypothesis is (necessarily) discredited. If the objective of an empirical research study is to discredit the null hypothesis (*i.e.*, to demonstrate or confirm "an effect", such as information content), a result just significant at 5%, when the sample size is very large, is usually not significant in any sense but the formal statistical. Moreover, on a more Bayesian analysis, the hypothesis tested may be strongly supported.

In orthodox terms, a very large test means that the null hypothesis is almost certain to be "rejected" at say α=5% even if the true parameter is very close to θ (assuming that the model is correct[20]); *i.e.*, power $\Pr\{\text{P-level} \leq \alpha \mid \theta_a\} \rightarrow 1$ for $\theta_a \rightarrow \theta_o$. It is no surprise, therefore, and not of great significance in itself, if, in a very large experiment, the observed level of significance is say .05; *cf.*, Berkson (1938, pp.526-527) and Hodges and Lehmann (1954, p.261). Moreover, if the sample size is sufficiently large, the orthodox

(two-sided) $100(1-\alpha)\%$ confidence interval corresponding to the result (two-sided P-level=α, n) is $[\theta_o, \theta_o+\varepsilon]$, where θ_o is the null (hypothesized) value of the unknown paramater θ, and ε (which may be negative) is arbitrarily small. Moreover, if the sample size is very large, a result significant at say $\alpha=.05$ (*eg.*, a P-level of say .04) may imply a confidence interval not including θ_o, but so close to θ_o and so short that if the true parameter θ is included in that interval, the null hypothesis, H_o: $\theta=\theta_o$, is as good as correct for any practical purpose.

In a case such as this, it is more to the point to report the $100(1-\alpha)\%$ confidence interval (for $\alpha=.05$ say) implied by the P-level observed than the P-level by itself. Generally, if the sample size is very large, a report which says that the observed P-level is equal to .049, or, more ambiguously, that the particular sample observed is significant at $\alpha=.05$, entails logically strong evidence of little or no real effect, but may not be interpreted as such by readers who routinely reject the (usually null) hypothesis of nil effect, or consider that hypothesis discredited, on seeing the locution "significant at $\alpha=.05$".

Appendix A

Table 2 below is a modification of the Dodd *et al.* Table 6 (p.24). The P-values shown are from the original table. For n greater than 19, these are calculated using the normal approximation of the binomial distribution (with continuity correction). Note that for n=36, the proportions of firms with negative abnormal returns given in the original table do not match the stated P-levels. The matching proportions are .56=20/36 (-1 to +1) and .53=19/20 (-2 to +2). Similarly, for n=19 the proportions of firms with negative returns implied by the P-levels stated are .63=12/19 (-1 to +1) and .58=11/19 (-2 to +2). In addition to these alterations, Table 2 includes the Bayesian posterior probability of H_O, $\Pr(H_O|x)$, calculated both from equations (1) and (2).

Table 2

Dodd *et al.* Table 6 Reconstructed

Type of Qualification	Window	n	x	$p=x/n$	P-level[a]	$\Pr(H_O\|x)$ [b]	$\Pr(H_O\|x)$ [c]
All	-1 to +1	283	165	.58	.0031	.1194	.0536
	-2 to +2	283	153	.54	.0954	.7435	.4961
Litigation	-1 to +1	130	76	.58	.0325	.4236	.2379
	-2 to +2	130	72	.55	.1271	.7085	.4739
Asset Realizing	-1 to +1	93	56	.60	.0307	.3652	.2130
	-2 to +2	93	50	.53	.2670	.7962	.5850
Multiple Uncertainties	-1 to +1	36	20	.56	.3089	.7255	.5282
	-2 to +2	36	19	.53	.4340	.7864	.5924
Future Financing	-1 to +1	19	12	.63	.1796	.5253	.3980
	-2 to +2	19	10	.53	.3238	.7498	.5684
Going Concern	-1 to +1	16	5	.31	.9616	.8876	.7350
	-2 to +2	16	8	.50	.5982	.7695	.5910
Disclaimers	-1 to +1	5	3	.60	.5000	.5882	.4912
	-2 to +2	5	3	.60	.5000	.5882	.4912

[a]Binomial test of H_O: $\theta = .5$ against H_A: $\theta > .5$.

[b]Based on the prior density function $g(\theta)$ constant over (.5, 1].

[c]Based on the prior density function $g(\theta) \propto \theta^2(1-\theta)^4$ over (.5, 1].

Note that the low posterior probability of the null hypothesis (-1 to +1) for the whole sample (for both priors) is attributable to firms in the litigation and asset realizing categories. For firms in the other categories, and for all firms over the wider observation interval, $Pr\{H_o|x\}$ is (for both priors) generally higher than $Pr\{H_o\}=.5$, indicating that the Dodd *et al.* binomial tests tend to confirm the null hypothesis that positive and negative abnormal returns are equally likely for firms with qualified audit reports.

Appendix B

Classical levels of significance should be interpreted in relation to the size of the sample. The importance of the sample size can be demonstrated from a Bayesian perspective by comparing the posterior probability of the hypothesis tested, $\Pr\{H_o|x\}$, conditioned on a fixed level of significance (P-level), over increasing values of the sample size n. Consider, for example, a (one-sided) test of significance on the null hypothesis $H_o: \theta=\theta_o=.5$ against $H_A: \theta>\theta_o$, given the model $x \sim$ Binomial (n,θ), where n is the sample size, θ is the probability of a "success" (θ is assumed constant from trial to trial), and x is the observed number of "successes" ($x \leq n$). This is the one-sided binomial test described by Dodd *et al.* Table 3 below shows the number and proportion of successes necessary to achieve given levels of significance .05, .10, and .25 for increasing values of n.

Table 3

Observed Proportion p Such That Level of Significance Equals P

n	P-level=.05		P-level=.10		P-level=.25	
	x[a]	$p=x/n$[b]	x	$p=x/n$	x	$p=x/n$
20	14	.7000	13	.6500	12	.6000
30	20	.6667	19	.6300	17	.5667
100	59	.5900	57	.5700	54	.5400
250	139	.5560	136	.5440	131	.5240
1000	527	.5270	521	.5210	511	.5110
2000	1037	.5185	1029	.5145	1016	.5080
10000	5083	.5083	5065	.5065	5034	.5034
100000	50261	.5026	50203	.5020	50107	.5011
1000000	500823	.5008	500641	.5006	500338	.5003
10000000	5002601	.5003	5002027	.5002	5001067	.5001
50000000	25005816	.5001	25004532	.5001	25002385	.5000

[a]The values of x given are those with P-levels closest to the stated levels. For n=20, the exact P-levels are .0577, .1316 and .2517. For larger n, the exact P-levels are very close to the levels stated.

[b]Rounded to four decimal places.

Note that as the sample size n increases, the proportion of successes $p=x/n$ with P-level equal to α approaches $\theta_o=.5$ (the higher α the closer $\{p$: P-level=$\alpha\}$ to θ_o). Intuitively, therefore, or in terms of the likelihood ratio, $\Pr\{x|H_o\}/\Pr\{x|H_a\}$, the higher n the wider the interval of alternative hypotheses $\theta \in [\theta_a, 1]$, where $\theta_a>\theta_o$, excluded or strongly refuted by a result just significant at α; *cf.*, Good (1981, p.164).

Hence, depending on the prior distribution, the more the Bayesian posterior distribution converges on H_O: $\theta=\theta_O$. For example, assume the "objective" (impartial) prior probability distribution $Pr\{H_O\}=Pr\{H_A\}=.5$, where $Pr\{H_A\}$ is distributed according to $g(\theta)$ uniformly over (.5,1]. Table 4 below gives the posterior probability of H_O, $Pr\{H_O|x\}$, given results with P-levels equal to .00001, .0001, .001, .01, .05, .10, .25, and .40 for increasing values of the sample size n.

Table 4

Bayesian $Pr\{H_O|x\}$[a]

n	P-level							
	.00001	.0001	.001	.01	.05	.10	.25	.40
20	.0000	.0002	.0113	.0464	.2877	.4616	.6094	.7157
30	.0000	.0004	.0085	.0783	.3102	.4601	.6934	.7664
100	.0002	.0026	.0226	.1855	.4540	.6231	.7881	.8501
250	.0006	.0045	.0377	.2622	.5791	.7242	.8596	.9009
1000	.0012	.0117	.0938	.4521	.7546	.8521	.9291	.9506
2000	.0018	.0151	.1172	.5205	.8269	.8951	.9477	.9660
10000	.0042	.0379	.2464	.7228	.9136	.9499	.9768	.9846
100000	.0137	.1090	.5140	.8946	.9714	.9840	.9926	.9951
1000000	.0427	.2829	.7702	.9641	.9909	.9949	.9976	.9984
10000000	.1239	.5555	.9141	.9884	.9971	.9984	.9993	.9995
50000000	.2405	.7365	.9597	.9948	.9987	.9993	.9997	.9998

[a]Based on the prior density function $g(\theta)$ constant over (.5, 1].

Note that as the sample size increases, the posterior probability of H_O, conditioned on $\{x$: P-level=$\alpha\}$ approaches 1 for fixed α. Moreover, if n is sufficiently large, a result just significant at .01, .05, .10 or any other level of significance, no matter how small, supports the null hypothesis. Thus, even if x is significant at $\alpha=.01$ or lower, if the sample size n is (very) large, the evidence supports the hypothesis tested. With regard to results "not significant" at conventional partitions such as $\alpha=.05$, if the observed level of significance is high, the sample size n need not be very large before H_O is much more strongly supported (more probable) than H_A. Thus, statistical insignificance may often represent extremely strong evidence *in favor* of the null hypothesis. This has been explained by both orthodox and Bayesian statistical theorists. References include Berkson (1942, p.331) Neyman (1955, pp.40-41), Gibbons and Pratt (1975, p.21), Birnbaum (1977, pp.37-38), Pratt (1977, pp.64-65) and Johnstone (1989e).

Footnotes

1. The framework on which Burgstahler's conclusion is deduced makes use of less than all available sample information. Johnstone (1989a) and Johnstone and Lindley (1989) suggest a more logical (Bayesian) framework.

2. *A priori*, a qualified audit report may make negative abnormal returns more probable but not less probable.

3. These are the terms used by Dodd *et al.* The "litigation" category includes firms involved in lawsuits of sufficient importance or potential consequence to warrant qualification by the auditors. The "asset realizing" group represents firms which, the auditors judge, have assets reported at figures above realizable value.

4. Evidence is measured conventionally by reference only to the level of significance (P-level) observed; there is no requirement that the sample size be taken into account (other than to calculate P). Nor is it necessary to consider the probability (or P-level) of the data on alternative hypotheses; *cf.*, Johnstone (1989b). See Seidenfeld (1979, pp.70-102) and Johnstone (1987a) for explanation of the classical (Fisherian) logic for statistical inference.

5. Hence, unless $g(\theta)$ is concentrated in the interval (.5,.58], the Bayes' factor in favor of H_O against H_A:

$$B = \frac{(.5)^n}{\int_{.5}^{1} \theta^x (1-\theta)^{n-x} \, g(\theta) \, d\theta}$$

where $g(\theta)$ represents the normalized prior density on H_A: $\theta > .5$, is greater for test (i) than test (ii).

6. For the purpose of comparing the relative force of the two tests, it is assumed that the prior probability distribution is the same for each test.

7. Note that if the sample sizes of the tests compared had been larger, the prior density on H_A: $\theta > .5$ would have to be concentrated in an even shorter interval (*i.e.*, $g(\theta)$ would need to be more perverse). For example, if the test sizes were n_1=1130 and n_2=1093 (adding 1000 to each), then, given the same P-levels as those actually observed, test (i) yields the lower posterior only if the density on $\theta > .5$ is almost all in the interval (.5,.527].

8. If we accept Bayes' theorem we are almost bound to use it. Inferences made other than in accord with Bayes' theorem (*e.g.*, the usual Fisherian inference, based on a low P-level, that the null hypothesis is discredited and probably false) are likely to imply, working backwards, an unacceptable prior distribution. A researcher who does not make explicit his prior belief distribution cannot draw an inference from empirical evidence without implying a prior probability distribution of some class. If priors of the class implied are not acceptable, then neither is the reseacher's stated inference, unless, of course, the laws of probability are ignored.

9. Thus no allowance is made for the possibility that θ is closer to .65 or .75 or higher. To a Bayesian this is dogmatic - the

Bayesian principle is to give very unlikely values of θ very small, but non-zero, prior probability. Otherwise, if the data strongly supports an improbable value of θ, the evidence will not be reflected in the posterior probability distribution; mathematically, a value of θ given zero prior weight gets zero posterior weight too, whatever the evidence.

10. The function *g* represents the normalized prior density for θ over $\theta > .5$. Thus, *g* can be interpreted as the probability (density) of θ given the condition that H_A: $\theta > .5$ is true.

11. Berger and Sellke (1987, pp.114-115) explain the use of priors with a "spike" of probability on θ_o. Briefly, if the hypothesis in question is not literally the point H_o: $\theta = \theta_o$ but an interval hypothesis h_o: $|\theta - \theta_o| \leq b$, *b* is often small enough (particularly if *n* is not extremely large) that $Pr\{h_o|x\}$ is approximated very closely by $Pr\{H_o|x\}$. Thus, the null hypothesis, written H_o: $\theta = \theta_o$, can be interpreted as an interval (disjunction) of point hypotheses, on which it is reasonable *a priori* to place non-zero probability.

12. Note, however, that for the purpose of demonstrating that the smaller test carries stronger evidence against H_o than the larger test, it is not necessary that the competing hypotheses, H_o and H_A, be given equal prior probability. It is necessary only that H_o have the same prior weight in each test. That is the respective starting points must be the same.

13. The incomplete beta function was computed using a power series outlined by Kennedy and Gentle (1980, pp.104-107).

14. This result is quite insensitive to the prior distribution on H_A: $\theta > .5$. Specifically, the posterior probability of H_o

conditioned on the observed value of x is low unless values of θ more strongly indicated by the data (*i.e.*, closer to the observed sample proportion p=.58) are ruled highly improbable *a priori*. For example, if the prior probability on H_A is distributed uniformly on the interval $.65 \leq \theta \leq 1$ say, $\Pr\{H_O|x\}$ equals:

$$(.5)^{283} / \left[(.5)^{283} + (1/.35) \int_{.65}^{1} \theta^{164}(1-\theta)^{119}\, d\theta\right] = .9514$$

But if the prior is distributed more evenly over all H_A: $\theta > .5$, the posterior probability of H_O is generally well below .5. For the prior represented by Figure 1 above, $\Pr\{H_O|x\}$=.0536.

15. It is often claimed that results which are not significant at conventional levels such as α=5% should be interpreted not as evidence in favor of the null hypothesis H_O, but merely as no (strong) evidence against H_O; *cf.*, Cox (1977, p.57; 1982, p.325), and Johnstone (1988, pp.322-323). This interpretation is not always appropriate; if the sample size is (very) large, middle and high P-values can entail extremely strong evidence in favor of the null hypothesis. See Appendix B.

16. This increase in the probability of H_O holds over a wide class of possible prior distributions. The probability of H_O is reduced only for prior distributions on H_A with a large concentration around θ=.54. But if the prior on H_A is distributed more evenly, or concentrated around values of θ greater than about .58, the P-level .0954 with n=283 represents evidence supporting H_O. Note that even for the prior density function $g(\theta)$ represented by Figure 1 in the text, which has over half its weight in the interval (.5,.6], the posterior probability of H_O decreases only marginally from .5 to .4961.

17. Data for firms filing 10-K reports with the SEC close to March 30 are considered relatively unreliable due to the great volume of reports processed by SEC personnel at about that time.

18. The relevance of the sample size is apparent from an orthodox standpoint if results are expressed as confidence intervals rather than P-levels. In general, the results (P-level=α, n_1) and (P-level=α, n_2), where $n_1 \neq n_2$, give rise to different confidence intervals at the same level of confidence (although the interval with the larger n is fully subsumed within the wider interval).

19. In most orthodox statistical textbooks, the result of a statistical hypothesis or significance test is expressed as either "accept H_o" or "reject H_o" at a fixed (predesignated) "critical" level of significance. This formal approach is due to Neyman and Pearson, and is widely accepted. For critical accounts of the Neyman-Pearson paradigm, see Seidenfeld (1979, pp.28-69) and Johnstone (1987b).

20. An invalid or inadequate model may reduce the power of a test [Lev and Ohlson (1982) p.270]. However, it is possible too that the power of the test (the probability of rejecting the null hypothesis conditional on a specified alternative) may increase; *cf.*, Johnstone (1989d). For example, if a relevant variable is excluded from a regression equation, the t-values of variables included (and correlated with the variable excluded) may be biased upward (or downward) towards the level required for rejection of the null.

References

BALL, R. and FOSTER, G. [1982]: "Corporate Financial Reporting: A Methodological Review of Empirical Research" (with discussion), *Journal of Accounting Research* 20, pp. 161-248.

BASU, D. [1975]: "Statistical Information and Likelihood" (with discussion), *Sankhya* A, 37, pp. 1-71.

BERGER, J.O. [1985]: *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. New York: Springer-Verlag.

BERGER, J.O. and DELAMPADY, M. [1987]: "Testing Precise Hypotheses", *Statistical Science* 2 (with discussion), pp. 317-352.

BERGER, J.O. and SELLKE, T. [1987]: "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence" (with discussion), *Journal of the American Statistical Association* 82, pp. 112-139.

BERKSON, J. [1938]: "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test", *Journal of the American Statistical Association* 33, pp. 526-542.

BERKSON, J. [1942]: "Tests of Significance Considered as Evidence", *Journal of the American Statistical Association* 37, pp. 325-335.

BIRNBAUM, A. [1977]: "The Neyman-Pearson Theory as Decision Theory, and as Inference Theory; With a Criticism of the Lindley-Savage Argument for Bayesian Theory", *Synthese* 36, pp. 19-49.

BURGSTAHLER, D. [1987]: "Inference from Empirical Research", *The Accounting Review* 62, pp. 203-214.

CASELLA, G. and BERGER, R.L. [1987]: "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem", *Journal of the American Statistical Association* 82, pp. 106-111.

COX, D.R. [1977]: "The Role of Significance Tests" (with discussion), *Scandinavian Journal of Statistics* 4, pp. 49-70.

COX, D.R. [1982]: "Statistical Significance Tests", *British Journal of Clinical Pharmacology* 14, pp. 325-331.

COX, D.R. and HINKLEY, D.V. [1974]: *Theoretical Statistics*. London: Chapman and Hall.

DeGROOT, M.H. [1986]: *Probability and Statistics*. 2nd ed. Reading, MA: Addison-Wesley.

DODD, P., DOPUCH, N., HOLTHAUSEN, R. and LEFTWICH, R. [1984]: "Qualified Audit Opinions and Stock Prices: Information Content, Announcement Dates, and Concurrent Disclosures", *Journal of Accounting and Economics* 6, pp. 3-38.

EDWARDS, W., LINDMAN, H. and SAVAGE, L.J. [1963]: "Bayesian Statistical Inference for Psychological Research", *Psychological Review* 70, pp. 193-242.

GIBBONS, J.D. and PRATT, J.W. [1975]: "P-values: Interpretation and Methodology", *The American Statistician* 29, pp. 20-25.

GOOD, I.J. [1980]: "The Diminishing Significance of a P-Value as the Sample Size Increases", *Journal of Statistical Computation and Simulation* 11, pp. 307-309.

GOOD, I.J. [1981]: "Some Logic and History of Hypothesis Testing", in *Philosophy in Economics*. (J.C. Pitt, ed.) pp. 149-174. Dordrecht: D. Reidel.

GOOD, I.J. [1982]: Discussion in Shafer, G. "Lindley's Paradox", *Journal of the American Statistical Association* 77, pp. 325-351.

GOOD, I.J. [1983]: "The Diminishing Significance of a Fixed P-Value as the Sample Size Increases: A Discrete Model", *Journal of Statistical Computation and Simulation* 16, pp. 312-314.

GOOD, I.J. [1984]: "The Tolerant Bayesian's Interpretation of a Tail-Area Probability", *Journal of Statistical Computation and Simulation* 19, pp. 300-302.

GOOD, I.J. [1985]: "Weight of Evidence: A Brief Survey" (with discussion), in *Bayesian Statistics 2*. (J.M. Bernardo *et al.*, eds.) pp. 249-269. Amsterdam: North Holland.

HILL, B.M. [1982]: Discussion in Shafer, G. "Lindley's Paradox", *Journal of the American Statistical Association* 77, pp. 325-351.

HODGES, J.L. and LEHMANN, E.L. [1954]: "Testing the Approximate Validity of Statistical Hypotheses", *Journal of the Royal Statistical Society* B, 16, pp.261-268.

JEFFREYS, H. [1939]: *Theory of Probability*. Oxford University Press.

JOHNSTONE, D.J. [1986]: "Tests of Significance in Theory and Practice" (with discussion), *The Statistician* 35, pp. 491-504.

JOHNSTONE, D.J. [1987a]: "Tests of Significance Following R.A. Fisher", *The British Journal for the Philosophy of Science* 38, pp. 481-499.

JOHNSTONE, D.J. [1987b]: "On the Interpretation of Hypothesis Tests Following Neyman and Pearson", in *Probability and Bayesian Statistics*. (R. Viertl, ed.) pp. 267-277. New York: Plenum.

JOHNSTONE, D.J. [1988]: "Comments on Oakes on the Foundations of Statistical Inference in the Social and Behavioral Sciences: The Market for Statistical Significance", *Psychological Reports* 63, pp. 319-331.

JOHNSTONE, D.J. [1989a]: "Inference from Empirical Research: A More Bayesian Approach", Working Paper #25, University of Sydney Accounting Research Centre.

JOHNSTONE, D.J. [1989b]: "Consistency and Relative Plausibility", *Journal of Statistical Computation and Simulation*, forthcoming.

JOHNSTONE, D.J. [1989c]: "Interpreting Statistical Significance When the Sample is Very Large", Working Paper, Graduate School of Management and Public Policy, University of Sydney.

JOHNSTONE, D.J. [1989d]: "Likelihood Logic in Orthodox Significance Tests", *Journal of Statistical Computation and Simulation*, forthcoming.

JOHNSTONE, D.J. [1989e]: "Interpreting Statistical Insignificance: A Bayesian Perpective", *Psychological Reports*, forthcoming.

JOHNSTONE, D.J. and LINDLEY, D.V. [1989]: "The Bayesian Inference Given Only that the Sample is Significant in Orthodox Terms", Working Paper, Graduate School of Management and Public Policy, University of Sydney.

JUDGE, G.G., HILL, R.C., GRIFFITHS, W.E., LUTKEPOHL, H., and LEE, T.C. [1982]: *Introduction to the Theory and Practice of Econometrics.* New York: Wiley.

KENDALL, M.G. and STUART, A. [1979]: *The Advanced Theory of Statistics*. Vol. 2, 4th ed. London: Griffin.

KENNEDY, W.J. and GENTLE, J.E. [1980]: *Statistical Computing.* New York: Marcel Dekker.

LEAMER, E.E. [1978]: *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

LEV, B. and OHLSON, J.A. [1982]: "Market Based Empirical Research in Accounting: A Review, Interpretation, and Extension" (with discussion), *Journal of Accounting Research* 20, pp. 249-331.

LINDLEY, D.V. [1957]: "A Statistical Paradox", *Biometrika* 44, pp. 187-192.

LINDLEY, D.V. and SCOTT, W.F. [1984]: *New Cambridge Elementary Statistical Tables*. Cambridge University Press.

NELSON, N., ROSENTHAL, R. and ROSNOW, R.L. [1986]: "Interpretaion of Significance Levels and Effect Sizes by Psychological Researchers", *American Psychologist* (November) pp. 1299-1301.

NEYMAN, J. [1955]: "The Problem of Inductive Inference", *Communications on Pure and Applied Mathematics* 8, pp.13-46.

PRATT, J.W. [1961]: "Review of Lehmann, E.L.: Testing Statistical Hypotheses", *Journal of the American Statistical Association* 56, pp. 163-167.

PRATT, J.W. [1976]: "A Discussion of the Question: For What Use Are Tests of Hypotheses and Tests of Significance" *Communication in Statistics -- Theory and Methods* A5, 8, pp. 779-787.

PRATT, J.W. [1977]: "'Decisions' as Statistical Evidence and Birnbaum's 'Confidence Concept'", *Synthese* 36, pp. 59-69.

ROSENKRANTZ, R.D. [1973]: "The Significance Test Controversy", *Synthese* 26, pp. 304-321.

ROSENKRANTZ, R.D. [1977]: *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*. Dordrecht: D. Reidel.

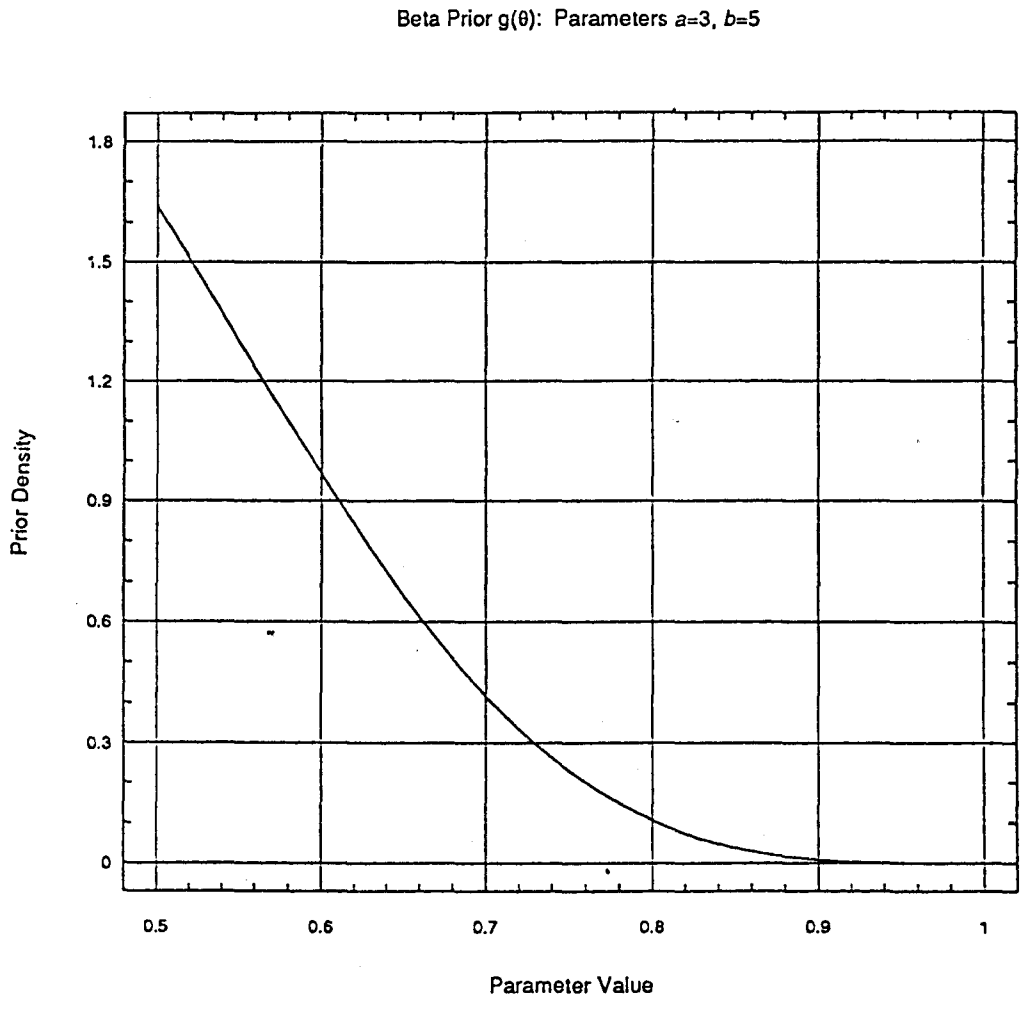
ROYALL, R.M. [1986]: "The Effect of Sample Size on the Meaning of Significance Tests", *The American Statistician* 40, pp. 313-315.

SEIDENFELD, T. [1979]: *Philosophical Problems of Statistical Inference: Learning From R. A. Fisher*. Dordrecht: D. Reidel.

ZELLNER, A. [1971]: *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

ZELLNER, A. [1984]: *Basic Issues in Econometrics*. University of Chicago Press.

Figure 1


Beta Prior g(θ): Parameters a=3, b=5
Prior Density
1.8
1.5
1.2
0.9
0.6
0.3
0
0.5
0.6
0.7
0.8
0.9
1
Parameter Value